Align Me : A framework to generate Parallel Corpus Using OCRs & Bilingual Dictionaries

Priyam Bakliwal **Devadath V V** **C V Jawahar**

CVIT, International Institute of Information Technology, Hyderabad, India

Abstract

Multilingual processing tasks like statistical machine translation and cross language information retrieval rely mainly on availability of accurate parallel corpora. Manual construction of such corpus can be extremely expensive and time consuming. In this paper we present a simple yet efficient method to generate huge amount of reasonably accurate parallel corpus with minimal user efforts. We utilize the availability of large number of English books and their corresponding translations in other languages to build parallel corpus. Optical Character Recognition systems are used to digitize such books. We propose a robust dictionary based parallel corpus generation system for alignment of multilingual text at different levels of granularity (sentence, paragraphs, etc). We show the performance of our proposed method on a manually aligned dataset of 300 Hindi-English sentences and 100 English-Malayalam sentences.

1 Introduction

Parallel corpus is an inevitable resource for many language processing tasks like Statistical Machine Translation(SMT) and cross-lingual information retrieval. Such tasks require an *aligned parallel corpus* where each sentence in a source language is aligned to the corresponding translated sentence(s) in target language. The task of creating a sentence aligned parallel corpus is expensive and time consuming since it involves the task of manual translation. Major sources for creating parallel corpus are Parliamentary proceedings like Europarl corpus(Koehn, 2005), parallel sentences from web and translations of books/documents.

India is a multilingual, linguistically dense and diverse country with rich resources of information (Chaudhury et al., 2008a). Though Monolingual corpora are available, availability of parallel corpus is very limited in quantity for language pair other than Hindi-English. Indian parliament proceedings are available only in Hindi and English and not in any other languages. But there are numerous amount of books that are translated in more than one language which are not digitized but can be used as a reliable source to generate parallel sentences. In this work, we are trying to leverage the Optical Character Recognition systems for digitizing the books in English and their respective translations in other Indian languages. For solving the problem of sentence alignment, various methods have been proposed over the past three decades like (Gale and Church, 1993). Since our data is OCR-generated data, existing algorithms failed to fetch a good level of accuracy since the text to be aligned is noisy.

To the best of our knowledge, two main algorithms have been proposed for sentence alignment in noisy data. The first work *Bleualign* (Sennrich and Volk, 2010) proposed MT based method for aligning sentences from OCR-generated parallel texts which are noisy. They used MT system to initially translate the texts and then used BLEU score(Papineni et al., 2002) to calculate the sentence similarity which is the base for alignment. Following this method, (Gomes, 2016) proposed a new scoring function that discriminates parallel and non-parallel sentences based on the ratio of text covered by bilingual phrase-pairs from a Moses phrase table. The first approach requires an MT system with a reasonable performance (Sennrich and Volk, 2010) which in our case is only possible for Hindi-English pair. The second method needs the access to bilingual-phrase pairs where for Indian languages have only limited number of sentences in the parallel corpus to create phrase tables.

The SMT systems are very sensitive towards the quality of training data. We have not come across any work in the past that have a mechanism to detect the failures of alignment algorithm. We propose an Active Learning based solution that does validations along with text alignment. The key idea is, if an algorithm is able to detect its failures and give that to a human in the form of queries, one can significantly reduce the amount of human effort while consistently maintaining the output quality.

In this paper, we propose a dictionary based recursive alignment algorithm to align text at multiple levels (sentence, paragraph, *etc.*). This method is a self updating validation algorithm that can predict when the alignment is

Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing,
pages 183–187, Osaka, Japan, December 11-17 2016.

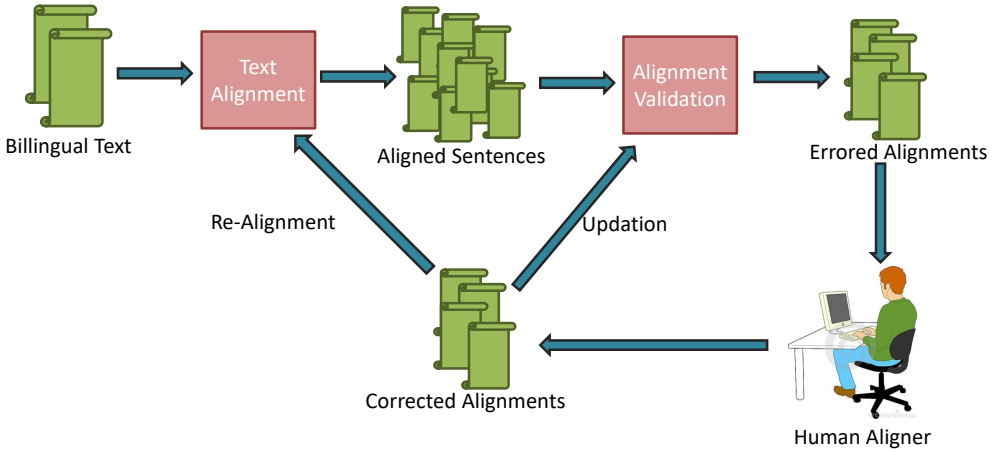

Figure 1: Block diagram of Align-Me framework. Given multilingual texts, an alignment algorithm is used to align the text. These aligned sentences are validated using length heuristics. Possible erroneous alignments are given to the user for corrections. These corrected alignments are used for updation of validation heuristics. In this way Align-Me aligns multilingual documents precisely with minimal user efforts.

done wrong. We show that the proposed framework can be used for precise alignment of multilingual sentences with minimal human effort.

2 Challenges for Data Creation & Sentence Alignment

These days the accuracy of OCR systems are very good. But still multiple errors occur while reading text due to font style difference, picture quality of book *etc.* Additional 1-to-many beads are introduced in our corpus by sentence boundaries being mis-recognized because of OCR or tokenization errors. There are several errors added in the form of spelling mistakes. Sentence alignment is further complicated by image captions, footnotes or advertisements that are not marked as such, and consequently considered part of the running text of the article. These text fragments typically occur at different positions in the two language versions, or only in one of them. They can be very disruptive to sentence alignment algorithms if they are not correctly recognized as deletions (1-to-0 or 0-to-1 beads), since a misalignment may cause consecutive sentences to be misaligned as well.

3 Algorithm

Align Me is an interactive framework that generates parallel corpus for two different languages given the parallel text (OCR data in our case) and a bilingual dictionary. As shown in Fig 1, the framework uses two separate algorithms: 'Alignment Algorithm' which align the sentences of the corpora and the 'Validation Algorithm' which detects where the former algorithm is failing. The sentences for which the alignment algorithm fails are given to the user for correction. Based on user corrections, the Validation algorithm updates itself for better prediction of the failures of the alignment algorithm.

We used the bilingual mappings released publicly by Indian Institute of Technology, Bombay (IIT, Bombay) for the initial alignment of text. These are dictionaries that contains root words of one language mapped to all its possible translationsin the other languages. There are 242 such dictionaries containing mappings of most of the Indian languages like Assamese, Bengali, Kannada, Gujarati, *etc.* Given the OCR generated parallel text T_{l1} and T_{l2} for language L_1 and L_2, we first find out all the words of language L_1 that occur exactly once in the T_{l1}. Further, We use a dictionary D_{l1-l2} to filter out the words from W_{l1} whose corresponding mapping in L_2 has occur only once. In this way we have a set of candidate aligned words C_{aw} in T_{l1} with their corresponding words in T_{l2}.

$$c_{aw} = \{(w_{l1}, w_{l2}) \mid \quad freq(w_{l1}) = freq(w_{l2}) = 1 \quad and \quad (w_{l1}, w_{l2}) \in D_{l1-l2}\} \tag{1}$$

It is observed that there exist a few erroneous items in word mappings found by Eq 1. Thus, we added another measure to validate the former mapping technique. We assume that the displacement of a word and its translation should not be large. We check that the relative position of two words w_{l1} and w_{l2} in the corresponding texts T_{l1} and T_{l1} should not differ more than a threshold τ.

$$f_{aw} = \{(w_{l1}, w_{l2}) \mid (w_{l1}, w_{l2}) \in C_{aw} \quad and \quad |(pos(w_{l1})/len(T_{l1}) - pos(w_{l2})/len(T_{l2}))| < \tau\} \tag{2}$$

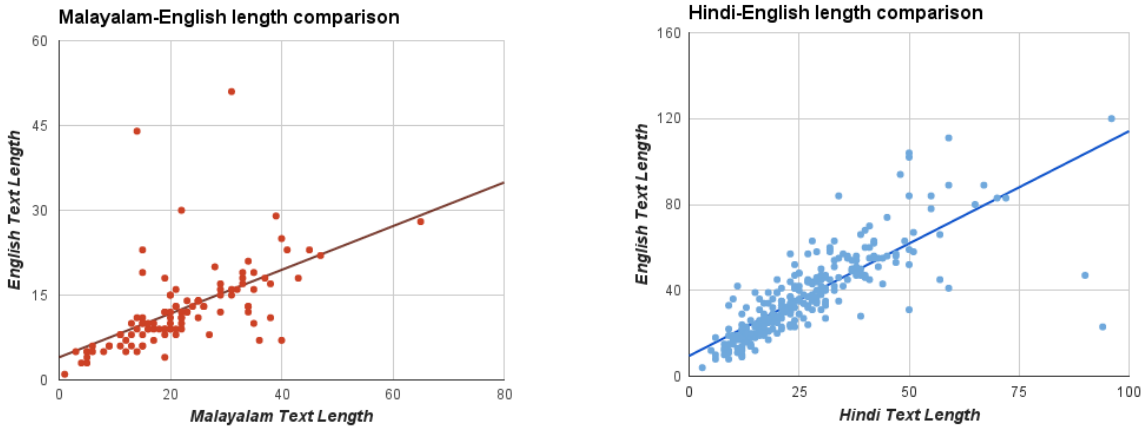

Figure 2: Comparison of number of words of 100 English-Malayalam sentences and 300 English-Hindi sentences. The figure shows that the count of words follow a nearly linear mapping.

where $pos(x)$ gives the position of a word in the text and $len(y)$ gives the length of the text. We consider the final word alignments f_{aw} as the correct alignments and use them as anchors to split the text. The next division of the text is done from the next separator. We use language specific sentence separators like "|", "?", "!" in Hindi and ".", "?", "!" in English.

Fig 2 shows that in spite of one-to-one or many-to-one mapping between sentences of two languages, the number of words in corresponding sentences mostly follow a linear mapping. This fact is used by our validation algorithm, we train a 'Linear Regressor' for the number of words present in the corresponding aligned texts of L_1 and L_2.

$$N_2 = a + b \times N_1 \tag{3}$$

where, N_1 and N_2 are number of words in aligned text of L_1 and L_2. We use the above trained Regressor to predict N_2 given N_1 for all the sentences aligned by the algorithm. The sentences where predicted number of words differs from that of original number of words by a certain threshold, are given to user for correction.

After the user corrections the Regressor is updated. These aligned texts are again given to the aligning algorithm for obtaining finer alignments. After each iteration we obtain finer annotations and an updated and more accurate Regressor.

4 Experiments & Results

To create the test data we digitized four books using OCR systems namely 'George Washington Man And Monument' and its Hindi translation and Kerala assembly Budget-speech of the year 2015 and its Malayalam translation. Due to the difference in writing styles of two authors, there is a huge difference between number of sentences present in the books and their respective translations. We have tested on 492 Hindi sentences and its corresponding 356 English sentences. We have aligned them manually to get 300 English-Hindi sentences. For English-Malayalam text we have used 140 Malayalam sentences and 165 English sentences. We created 100 English-Malayalam aligned sentences to validate the performance of proposed approach.

The approaches proposed in the past used various evaluation measures. Dan (1996) used block error to evaluate alignments. Chaudhary *et. al* (2008b) proposed a sentence based evaluation using Precision, Recall and F1-Measure. For the first level alignment of Hindi-English text we are getting 85.2% precision and 78% recall and for Malayalam-English text we are getting 96% precision and 85% recall.

To show the effectiveness of 'Active Learning' in the alignment task, we have used 'Word Level Error' than 'Sentence Level Error'. Even if a single word of a sentence have a mis-alignment, all the other words of that sentence are said to be aligned erroneously. We calculate 'Word Error Percentage' for both the languages as $(NumberofMisalignedWords/TotalNumberofWords)$. In Fig 4 we show that our algorithm is able to detect correctly, the mis-aligned texts to be queried to the user. The figure shows the reduction in error with every user correction for two iterations on same text.

Fig 3 shows that Align-Me is effectively able to detect aligned texts of different modularities. With each iteration finer alignments are done. We also show that the proposed framework is immune to OCR system introduced errors. In the second iteration of Malayalam alignment, the algorithm handled 1-to-many beads introduced due to misrecognition of sentence boundaries by OCR systems.

Figure 3: The above table shows the qualitative performance of Align-Me. The top row depicts the output of first iteration and bottom row depicts the output of second iteration. One can get aligned sentences at different levels depending on the requirement.

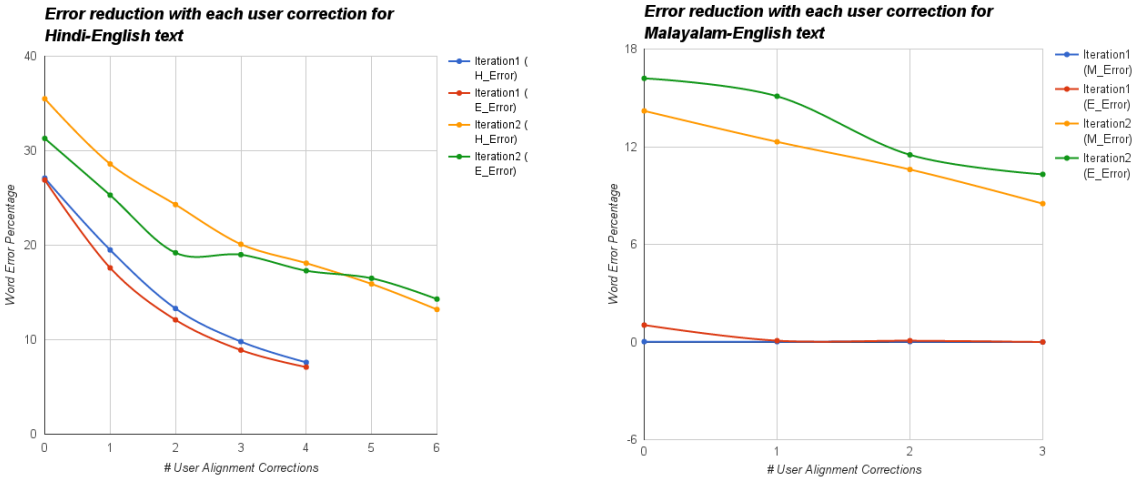

Figure 4: The above graph shows the reduction of 'Word Error Percentage' with every user annotation. We have calculated word errors for all the languages. 'H_Error', 'E_Error' and 'M_Error' are word errors for Hindi, English and Malayalam respectively. The error graph shows the fall of error for two iterations. It is evident that the validation algorithm is able to correctly determine the mis-aligned samples.

5 Conclusion & Future Directions

In this paper, we proposed Align Me as an efficient framework for generating huge corpus of parallel text using minimal user efforts. Our framework uses multilingual dictionaries to align the texts initially. At every step, the verification of the alignments is done using a validation algorithm which uses length based heuristics to determine possible mis-alignments. Experimental data depicts that length based heuristics work really well in cases where there are possible errors in the text-to-be-aligned. These heuristics also perform exceedingly well in cases when the number of sentences in both the languages vary by a huge count. In this approach, the human effort is reduced to a great extent as the framework queries only the misaligned sentences to the human annotator. The proposed approach can be utilized for generation of huge corpus for languages like Malayalam-English, Marathi-English, Hindi-Kannada *etc.* where there is huge paucity of aligned data. The performance of the method remains consistent even if the input data is noisy; this proves the high degree of robustness that the method offers.

As part of future work, we would like to use the proposed framework for generation of parallel corpus for other Indian languages as well. We are also trying to incorporate other factors like BLEU score for detection of mis-alignments.

References

Sriram Chaudhury, Dipti Misra Sharma, and Amba P Kulkarni. 2008a. Enhancing effectiveness of sentence alignment in parallel corpora: Using mt heuristics.

Sriram Chaudhury, Dipti Misra Sharma, and Amba P Kulkarni. 2008b. Enhancing effectiveness of sentence alignment in parallel corpora: Using mt heuristics. *ICON*.

William A Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Luıs Gomes. 2016. First steps towards coverage-based document alignment.

IIT. Bombay. Bilingual mappings (http://www.cfilt.iitb.ac.in/downloads.html).

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

I Dan Melamed. 1996. A geometric approach to mapping bitext correspondence. *arXiv preprint cmp-lg/9609009*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. Mt-based sentence alignment for ocr-generated parallel texts.